IMPROVING THE ACCURACY OF PREDICTING LUNG CANCER PATIENT SURVIVAL TIME USING LSTM NEURAL NETWORK

Nazeer K.A.A.[1], Sumanathilaka T.G.D.K.[2*], Toppo E.R.[1], and Lakra R.[1]
[1]Department of Computer Science and Engineering, National Institute of Technology Calicut, India
[2]Department of Computing, SLIIT, Sri Lanka
*deshan.s@sliit.lk

Outcomes for cancer patients have been previously estimated by applying various machine learning techniques to large data-sets such as the Surveillance, Epidemiology, and End Results (SEER) program database. In particular for lung cancer, it is not well understood which types of techniques would yield more predictive information, and which data attributes should be used in order to determine this information. In this study, a number of supervised learning techniques are applied to the SEER database to classify lung cancer patients in terms of survival, including Multi-Layer Perception (MLP), Long Short Term Memory (LSTM) and Deep belief model (DBM). Key data attributes in applying these methods include tumor grade, tumor size, gender, age, stage, and the number of primaries, with the goal to enable comparison of predictive power between the various methods. The prediction is treated like a continuous target, rather than a classification into categories as the first step towards improving survival prediction. Several types of research have been performed on the current topic and the best result was achieved using Custom ensemble with RMSE value 15.30. The LSTM Neural Network was trained and tested in order to improve the accuracy of the predication and the below results were achieved. The dataset with more than a hundred thousand instances on lung cancer patients' details was downloaded from the SEER training model, National Cancer Institute and principal components for training were identified using Principal Component Analysis (PCA). The data was divided into 75% as training data and rest for testing. The best performing technique was Long Short Memory (LSTM) with a Root Mean Square Error (RMSE) value of 10.53 months.10 fold cross-validation was performed over the dataset and results were taken. This results proven that LSTM Neural Network can be successfully used for estimating patient survival time with the ultimate goal to inform patient care decisions.

Keywords: *Lung Cancer, SEER dataset, Long short term memory, Supervised Machine Learning, Multi-layer Perception*